"Phrases and associations across languages: experiments in corpus-based computational phraseology"

Colson, Jean-Pierre

Abstract

The automatic extraction of all collocations / phraseologisms from corpora has a crucial role to play in the development of computational phraseology. Unfortunately, after « 50-something years of work on collocations », the results are still disappointing (Gries 2013). A possible way of improving the results is to start, not from traditional statistical scores, but from other techniques inspired from information retrieval, and in particular metric clusters. The Corpus Proximity Ratio (CPR, J.-P. Colson 2014) makes it possible to reach a precision level of about 85 percent for the extraction of bigrams, but also of higher grams (up to sevengrams). In this paper we show the results obtained by a combination of raw frequency (on web corpora of 200 million words) and of CPR for the English formula It's not and its French counterpart C'est pas, as well as for the French pattern en toute and its counterparts in English, German and Spanish (resp. in all / in aller / en toda). The results ...

Document type : *Communication à un colloque (Conference Paper)*

Abstract

We refer here to phraseology as the study of set phrases in the broadest sense, including partly fixed phrases (routines and formulae, collocations), and also very fixed phrases (idioms and proverbs). Contrary to other studies centered around collocations, computational phraseology (Heid 2007, Corpas Pastor 2013) investigates the close relationship between all categories of set phrases or phraseologisms, presents them as a continuum ranging from low to high idiomaticity and/or fixedness, and considers set phrases as one of the foundation stones of linguistic theory.

Corpus-based computational phraseology has recourse to huge corpora, as it has been demonstrated that, while set phrases as a whole represent a high percentage of word combinations in any text (Sinclair 1991), the frequency of a given set phrase in a corpus is usually low (Colson 2007, Moon 1998). Computational phraseology largely relies on the statistics of word co-occurrence (Evert 2004), but also explores other techniques derived from information retrieval, such as clustering (Baeza-Yates & Ribeiro-Neto 1999).

The automatic extraction of all collocations / phraseologisms from corpora has a crucial role to play in the development of computational phraseology
Unfortunately, after « 50-something years of work on collocations », the results are still disappointing (Gries 2013). A possible way of improving the results is to start, not from traditional statistical scores, but from other techniques inspired from information retrieval, and in particular metric clusters. The *Corpus Proximity Ratio* (CPR, J.-P. Colson 2014) makes it possible to reach a precision level of about 85 percent for the extraction of bigrams, but also of higher grams (up to sevengrams). In this paper we show the results obtained by a combination of raw frequency (on web corpora of 200 million words) and of *CPR* for the English formula *It's not* and its French counterpart *C'est pas,* as well as for the French pattern *en toute* and its counterparts in English, German and Spanish (resp. *in all* / *in aller / en toda*). The results show that such structures encompass a wide variety of communicative and semantic phrases, most of which are nowhere to be found in dictionaries. With the help of *CPR,* no less than 150 phrases were extracted with *C'est pas* at the beginning of the phrase.